Simulating Success: GPT-4's Impact on the Development of a Virtual Communication Training Skills Module for Medical Students

Dan Weisman, Alanna Sugarman, Yue-Ming Huang, Lillian Gelberg, Patricia A. Ganz, Warren Scott Comulada

Submitted to: JMIR Medical Education
on: August 27, 2024

Table of Contents

Simulating Success: GPT-4's Impact on the Development of a Virtual Communication Training Skills Module for Medical Students

Dan Weisman[1] MFA; Alanna Sugarman[2] BSc; Yue-Ming Huang[3, 1] FSSH, MHS, EdD; Lillian Gelberg[4, 5] MSPH, MD; Patricia A. Ganz[2, 5] MD; Warren Scott Comulada[5, 6] DrPH

[1]UCLA Simulation Center University of California, Los Angeles Los Angeles US
[2]David Geffen School of Medicine University of California, Los Angeles Los Angeles US
[3]Department of Anesthesiology and Perioperative Medicine University of California, Los Angeles Los Angeles US
[4]Department of Family Medicine University of California, Los Angeles Los Angeles US
[5]Department of Health Policy and Management University of California, Los Angeles Los Angeles US
[6]Department of Psychiatry and Biobehavioral Sciences University of California, Los Angeles Los Angeles US

Corresponding Author:
Warren Scott Comulada DrPH
Department of Psychiatry and Biobehavioral Sciences
University of California, Los Angeles
760 Westwood Plaza, 37-384C
Los Angeles
US

Abstract

Background: Standardized patients (SPs) prepare medical students (MSs) for difficult conversations with patients, such as discussions about life-changing diagnostic results. Despite their value, SP training is constrained by available resources and competing clinical demands. Researchers are turning to generative pre-trained transformers (GPTs) and other large language models (LLMs) to create communication skills simulations that incorporate computer-generated (virtual) SPs (VSPs). GPT-4 is a major LLM advance that makes it practical for developers to use text-based prompts instead of Branching Path Simulations (BPS) that rely on pre-scripted conversations. These nascent developmental practices have yet to take root in the literature to guide other researchers in developing their own simulations.

Objective: This study aims to describe our developmental process and lessons learned for a GPT-4-driven VSP. We designed the VSP to help MS learners rehearse discussing abnormal mammography results with a patient as a primary care physician (PCP).

Methods: We conducted in-depth interviews with 5 MSs, 5 PCPs, and 5 breast cancer survivors to inform development of the scenario and VSP. We then used Hyperskill, simulation authoring software, to develop a VSP. Initially, GPT-4 was not available. We started development using BPS. Aided by GPT-4, we used a prompt to instruct the VSP regarding the scenario, its emotional state, and expectations for how the learner should converse with it. We iteratively refined the prompt after multiple rounds of testing. As an exploratory feature, we programmed the simulation to display written feedback on the learner's performance in communicating with the VSP.

Results: In-depth interviews helped us create a realistic scenario by establishing when a conversation between a PCP and patient would likely first take place in the breast cancer screening process and the mode of communication. The scenario simulates a telephone call between the learner and patient to discuss the results of an abnormal diagnostic mammogram that requires a biopsy. Interviews informed programming of prompts for the VSP to expect learner communication based on the SPIKES protocol for delivering bad news. The simulation also evaluated the learner's performance based on the SPIKES protocol. Preliminary testing was promising. The VSP asked sensible questions about their mammography results and responded to learner inquiries using a realistic voice replete with appropriate emotional inflections based on the conversation. Feedback was useful to highlight major SPIKES deviations but less so when clinical judgement was warranted to balance VSP responses with appropriate next steps (e.g., not pressuring the VSP to schedule a biopsy while displaying agitation).

Conclusions: GPT-4 streamlined development and provided a better and more natural user experience than what we were able to provide using BPS. As next steps, we will continue to develop the simulation to improve feedback and pilot test the VSP with MSs to evaluate its feasibility.

(JMIR Preprints 27/08/2024:65670)
DOI: https://doi.org/10.2196/preprints.65670

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Title: Simulating Success: GPT-4's Impact on the Development of a Virtual Communication Training Skills Module for Medical Students

Abstract

Background: Standardized patients (SPs) prepare medical students (MSs) for difficult conversations with patients, such as discussions about life-changing diagnostic results. Despite their value, SP training is constrained by available resources and competing clinical demands. Researchers are turning to generative pre-trained transformers (GPTs) and other large language models (LLMs) to create communication skills simulations that incorporate computer-generated (virtual) SPs (VSPs). GPT-4 is a major LLM advance that makes it practical for developers to use text-based prompts instead of Branching Path Simulations (BPS) that rely on pre-scripted conversations. These nascent developmental practices have yet to take root in the literature to guide other researchers in developing their own simulations.

Objective: This study aims to describe our developmental process and lessons learned for a GPT-4-driven VSP. We designed the VSP to help MS learners rehearse discussing abnormal mammography results with a patient as a primary care physician (PCP).

Methods: We conducted in-depth interviews with 5 MSs, 5 PCPs, and 5 breast cancer survivors to inform development of the scenario and VSP. We then used Hyperskill, simulation authoring software, to develop a VSP. Initially, GPT-4 was not available. We started development using BPS. Aided by GPT-4, we used a prompt to instruct the VSP regarding the scenario, its emotional state, and expectations for how the learner should converse with it. We iteratively refined the prompt after multiple rounds of testing. As an exploratory feature, we programmed the simulation to display written feedback on the learner's performance in communicating with the VSP.

Results: In-depth interviews helped us create a realistic scenario by establishing when a conversation between a PCP and patient would likely first take place in the breast cancer screening process and the mode of communication. The scenario simulates a telephone call between the learner and patient to discuss the results of an abnormal diagnostic mammogram that requires a biopsy. Interviews informed programming of prompts for the VSP to expect learner communication based on the SPIKES protocol for delivering bad news. The simulation also evaluated the learner's performance based on the SPIKES protocol. Preliminary testing was promising. The VSP asked sensible questions about their mammography results and responded to learner inquiries using a realistic voice replete with appropriate emotional inflections based on the conversation. Feedback was useful to highlight major SPIKES deviations but less so when clinical judgement was warranted to balance VSP responses with appropriate next steps (e.g., not pressuring the VSP to schedule a biopsy while displaying agitation).

Conclusions: GPT-4 streamlined development and provided a better and more natural user experience than what we were able to provide using BPS. As next steps, we will continue to develop the simulation to improve feedback and pilot test the VSP with MSs to evaluate its feasibility.

Key Words: Standardized patient, LLM, GPT-4, agent, abnormal mammography, biopsy

INTRODUCTION

OpenAI's Generative Pre-Trained Transformer 4 (GPT-4) and other large language models (LLMs) that use advanced artificial intelligence (AI) algorithms to mimic human responses to text and voice queries are rapidly changing medical education and practice. Medical students and residents are using LLMs to paraphrase complex medical concepts for easier understanding, create self-study

questions for medical exams, summarize research papers and generate written e-mail responses, among other tasks [1]. GPT-4 has the potential to generate questions for medical school exams, help grade them, and provide written feedback to students [2]. Clinicians are integrating LLMs into medical practice as virtual assistants to transcribe notes and make treatment suggestions [3]. LLMs also interact with patients as conversational agents (i.e. chatbots) to book appointments, manage medical records, and draft responses to patient questions [3,4]; one study found LLM-generated responses to be preferable by patients, likely due to the ability of LLMs to eloquently respond without the workload of a human to hinder a similar level of response quality [5].

Chatbots designed to train clinicians to converse with patients during clinical visits are a nascent LLM application that is the focus of this paper. Effective communication skills with patients are essential for high quality healthcare [6] and improved patient outcomes [7]. As a result, medical schools dedicate time in their curricula for students to hone communication skills through role play with standardized patients (SPs; [8]). However, learning communication skills training, especially related to serious illnesses, is often overshadowed by competing clinical training demands [9]. Additionally, selecting and training SPs to portray the authenticity of actual patients is challenging [8,10]. In response, researchers are developing chatbots to act as virtual standardized patients (VSPs) that use LLMs to simulate conversations with learners [11-14]. For example, Holderried et al. developed a GPT-3.5-powered chatbot for medical students to practice taking patient histories [11].

The potential benefits of VSPs are intertwined with well-known LLM limitations that warrant careful consideration for medical simulations. The flexibility of LLMs to deviate from scripted conversations offers more human-like interactions but can lead to too much improvisation, creating off-topic or biased responses from inappropriate source data and fabricating information when no source data is available, referred to as “hallucination” [15-17]. Additionally, the complexity of conversations between physicians and patients can push LLMs to their limit in trying to mimic patients [18] and providing feedback to learners about their performance conversing with VSP. Feedback is an important aspect of SP training to improve learners’ communication skills [19] and is a difficult task, even for human evaluators [10,20]. This is a reason VSP-based training scenarios have had limited evaluation capabilities, at best providing point-based evaluations [21].

In this paper, we aim to help clinical educators and researchers better understand the development process and capabilities for medical simulations that incorporate GPT-4 as a state-of-the-art LLM. We do this in the context of a GPT-4-based VSP we developed for a medical simulation study. The learner, taking on the role of a primary care physician (PCP), simulates a phone call with the VSP to discuss an abnormal screening mammogram result that requires a patient to schedule a biopsy. Our work contributes to burgeoning literature on the architecture of LLM-based medical communication skills training modules. Existing literature presents general frameworks for LLM-based simulations across disparate clinical scenarios with a focus on clinical reasoning and diagnosis [21, 22]. We concentrate on a single scenario to allow for an in-depth discussion of the simulation development process and LLM integration that is applicable to a plethora of training scenarios anchored in the uncertainties of screening tests. We first present qualitative work that informed the development process and then discuss a standard process for developing communication training skills simulations. Next, we discuss how we pivoted to streamline our development process by capitalizing on GPT-4 as a major LLM advancement in its ability to incorporate clinical context into its dialogue [23-26]. Lastly, we discuss exploratory work to automatically generate AI feedback on the learner’s performance during their simulated conversation with the VSP.

METHODS

Overview

We designed a simulation scenario for medical students to practice discussing with a patient the results of an abnormal mammogram that requires a patient to schedule a biopsy. The simulation

portrays a realistic conversation between a learner (role-playing as a PCP) and a VSP as a first step to develop a suite of communication skills training simulations for medical students to practice delivering difficult news to patients. We started with the scenario of an abnormal mammogram because of the uncertainty and anxiety it can invoke in a patient. Breast cancer is the most common incident cancer among women in the US and worldwide [27]. Yet abnormal mammograms that require biopsy may not reveal cancer because mammography does not have high specificity that require follow-up [28]. This study consisted of two phases: 1) a formative phase with in-depth stakeholder interviews that guided design decisions and 2) a development phase to create the virtual training module. The study took place from September 2023 to August 2024. We obtained ethical approval for all study procedures from the Institutional Review Board at [blinded for review].

Formative phase

We conducted 30- to 60-minute in-depth interviews with medical students at our institution (MSs; n=5), PCPs (n=5), and breast cancer survivors (BCSs; n=5) who had received a breast cancer diagnosis within the past five years as key stakeholders. Recruitment occurred through referrals and word-of-mouth. Interested individuals filled out electronic (Qualtrics) screening and consent forms to enroll in the study. See Appendix for interview guide questions.

We audio-recorded and transcribed interviews for thematic analysis. We analyzed transcripts in two steps. First, we used Microsoft Copilot, a LLM chatbot like ChatGPT, to identify themes from the transcript text. Similar to how users interact with ChatGPT, we typed instructions (a prompt) for Copilot as to the format of the transcripts, provided relevant details about each stakeholder group and an overview as to the purpose of the interviews (e.g., “medical students” and “their experience in medical school receiving training to deliver bad news and if they had experience delivering bad news to patients”), instructed Copilot to play the role of a “qualitative researcher” and to “summarize the content of the interview transcripts and come up with common themes across the interviews”. Second, co-authors who conducted interviews reviewed Copilot-generated summaries for each stakeholder group and finalized themes after consensus among the research team. Tables 2 and 3 present themes and illustrative quotations. Quotations from MSs are indicated by “MS” followed by their year of medical school and a lowercase letter, starting with “a”, if there is more than one MS for a given year. For example, MS4a indicates the first fourth year medical student we interviewed. PCP and BCS numbers indicate the order in which they were interviewed, e.g., PCP1 indicates the first PCP we interviewed.

Development phase

Software authoring tool

We used Hyperskill software (SimInsights, Lake Forest, CA) to develop the VSP [29]. Hyperskill features an authoring interface for developing custom learning experiences and uses Automated Speech Recognition (ASR) and Text-to-Speech (TTS) technology to facilitate real-time simulated conversations that users conduct using natural speech. Beta features integrating generative AI and LLMs allowed us to pivot our simulation design from a rigidly scripted branching scenario to a more open-ended and naturalistic conversation driven by role-playing AI “prompts” as described below.

Original design: Scripted branching scenario

We designed the original scenario for the simulation as a Branching Path Simulation (BPS) based on the structured Setting, Perception, Invitation, Knowledge, Emotion, and Summarize (SPIKES) protocol for delivering bad news [30,31]. The formative interviews drove our decision to use SPIKES as discussed in the Results. We wrote scripted dialogue for both the learner and the VSP and created a scenario flowchart, with each step of SPIKES represented as a new stage on the flowchart.

As illustrated in Figure 1, the design of a BPS scenario for a difficult phone call with a VSP involves mapping out a complex web of potential responses and outcomes, with numerous critical decision points. We wrote scripted dialog encompassing a wide range of possible positive and negative interactions between the user and the VSP. The large number of branch points introduced substantial obstacles to authoring our prototype within existing time and budget constraints, and still did not adequately cover the numerous possible outcomes of a real-life patient conversation.

As part of the original BPS design, we evaluated a scoring system called voice intent classification. This system compares spoken user responses to a list of fixed utterances, provides a score based on how closely they match, and chooses where to branch the scenario based on this score. At each critical decision point in our BPS training, we prepared a list of five scripted examples of ideal, acceptable, and unacceptable user responses which were assigned scores of 1, 0.5, and 0 points respectively. The system would then trigger the appropriate scripted responses for the VSP based on the score of each user response.

Testing showed that this scoring system did not provide an accurate or valid assessment of user communication abilities because the system consistently failed to recognize matches for many correct variations of user responses. When the system failed to recognize a match, a multiple-choice question was triggered as a fallback and users would pick their preferred response from the pre-written list of utterances. This limited the effectiveness of the training because ideal responses were much easier to identify when presented alongside clearly unacceptable responses.

Modified design using LLM integration

After encountering challenges with authoring a BPS scenario, we explored the latest advances in LLM role-playing abilities to create a more dynamic scenario. We modified our design using a beta feature that provided integration with a GPT-4-powered chatbot. This feature eliminated the need for us to script dialog or design complex branching scenarios by enabling the VSP to converse with the learner using completely AI generated responses. This simplified the scenario and streamlined the development process of the prototype, conserving both time and resources. In contrast to the six-stage, multi-path, scripted dialog-heavy design of the BPS scenario, the AI-driven VSP scenario design is dictated by a text-based set of instructions (a prompt) that guides chatbot responses. The final scenario flow used for the simulation consists of only two states: Setup and Feedback (see Figure 2), whereas a BPS scenario would contain dozens of additional states for each scripted line of dialog and critical decision point. Testing showed that the detailed text prompt we provided for the VSP chatbot produced similar branching conversation paths as our planned BPS scenario.

For post-simulation feedback, we incorporated another beta feature that allowed us to add a GPT-4 enabled “AI agent” to the simulation. AI agents are advanced systems that autonomously interact within digital environments, make decisions, and perform actions based on the language understanding provided by a LLM [32]. Like chatbots, AI feedback agents can be guided by a prompt which provides them with a pre-defined role and specific instructions. Unlike a chatbot, they can then execute actions based on this prompt, such as analyzing the entire content of simulated conversations, identifying good and bad examples of user quotes, and implementing a point-based scoring system. For this simulation, we created an AI agent that analyzes the user’s entire conversation with the VSP and gives automated feedback on the learner’s performance after their simulated conversation ends.

RESULTS

Formative work

Sample characteristics

Table 1 displays sociodemographic background characteristics of in-depth interview participants. MSs and PCPs were diverse in terms of sex at birth, and race/ethnicity. MSs were represented across the second, third, and fourth years of medical school with most MSs in their fourth year. There was less ethnic and racial diversity among BCSs with four BCSs identifying as non-Hispanic White and only one identifying as Hispanic. BCSs varied in age from 38 to 76 years old. Next, we summarize in-depth interview findings by themes.

Need for additional communication skills training

Interviews with all three stakeholder groups validated findings in the literature that call for additional communication skills training beyond what medical schools and clinical settings have the capacity to provide. MSs discussed three forms of communication skills training they receive: didactic training that provides protocols for delivering difficult news like “SPIKES”, SP training, and clinical rotations where they interact with actual patients and receive guidance from attending physicians. MSs valued training, including SP sessions where students practiced delivering bad news to SPs (e.g., “a cancer diagnosis” and “STI diagnosis”), and received feedback to improve their communication skills as indicated by MS quotations in Table 2. MSs also indicated that communication skills training was limited in medical school. SP interactions took place in groups. Students rotated in and out of the role as learners where they acted as physicians conversing with SPs while other students observed. Therefore, not all the students had a chance to practice delivering bad news to SPs. One MS indicated the limitation of SPs to reflect the range of patient reactions and emotions, since SPs followed a script.

MS indicated how clinical rotations helped them put SP training into practice under the supervision of attending physicians. They also learned by observing their attending physicians talk to patients. For example, one MS commented on the cultural sensitivity they observed (Table 2). MSs also indicated limitations honing their communication skills during rotations. Practice delivering the news of serious diagnoses to patients varied across clinical settings. One student assigned to a “County hospital site” had a lot of experience due to the volume of cases and limited capacity of attending physicians to oversee them. Similar to MSs, PCPs recalled protocols for delivering bad news, such as “SPIKES” [30,31] and “ABCDE” [33], but they reinforced the notion that classroom preparation was limited in its ability to prepare MSs to engage in difficult conversations with patients.

Openness to new types of communication skills training

The context of our interviews to inform the development of a VSP gave us opportunities to ask students about their exposure to virtual training and interest in 2D versus virtual reality options. MSs had limited exposure but recognized its potential to improve communication skills training (Table 2).

Patient journey through the breast cancer screening process

PCP and BCS interviews helped us understand when PCPs engage with patients in breast cancer screening and diagnostic processes and how PCPs typically communicate with patients (i.e., through the telephone, a video conference call, or in-person visit) to develop a realistic training scenario.

PCPs reported limited patient contact during the early stages of the breast cancer screening process (Table 2). Electronic health record systems send automated messages to patients that prompt them to schedule screenings, inform them about screening test results, and prompt patients to schedule additional follow-up exams when needed. PCPs talk to patients if additional examination is needed beyond a screening mammogram when a cancer diagnosis is more likely and patient concern is higher, e.g., for a biopsy.

We also discovered that PCPs typically communicate with patients via a message in the EHR patient portal or telephone call during the breast cancer screening process. PCPs schedule in-person or video visits to deliver more serious news to the patient, such as a mammogram with a higher probability of a cancer diagnosis.

Interview discussions also shaped how we prompted the VSP to respond to learners. Given the emphasis on the SPIKES protocol by MSs and PCPs as a guide for presenting difficult news to patients and the alignment of communication-related interview themes with SPIKES (Table 3), we prompted our VSP chatbot to expect communication based on the SPIKES protocol.

Final design specifications for the virtual simulation

We designed the simulation to portray a scenario where a patient has had a diagnostic mammogram due to an abnormal screening mammogram. The patient now needs to schedule a biopsy based on a suspicious mass revealed by the diagnostic mammogram. They spoke with a radiologist about the results of their diagnostic mammogram and a biopsy was recommended for further testing and diagnosis. The learner, taking on the role of the patient's PCP, is tasked with discussing the results of the diagnostic mammogram (the patient has already seen the mammogram results in their patient portal) and encouraging the VSP to proceed with the next step of treatment (scheduling a biopsy) through a simulated telephone call. The VSP displays behaviors consistent with common feelings of anxiety, worry, and uncertainty described by BCS interviewees who had similar discussions with their physicians.

User interface and virtual environment

For the environment of the virtual simulation, we used a simple three-dimensional (3D) virtual object depicting a cell phone laying on a doctor's office desk (see Figure 3). Our chosen scenario of a simulated phone call represented the typical mode that PCPs reported using for discussing with a patient results of a diagnostic mammogram with a suspicious mass, and a phone call eliminated the need for building an immersive 3D environment or implementing animated character models, instead allowing us to focus on creating a screen-based chat focused training with relatively minimal technical requirements.

We implemented an "End" button on the virtual cell phone object displayed on screen. Users are instructed to click the red "End" button when the conversation with the VSP has concluded. When clicked, this button simulates hanging up the phone, triggers the end of the scenario, and begins generating feedback based on the user's performance.

Initial VSP prompt construction

The goal of this prompt was to generate a realistic, flexible, and consistent simulated patient who interacts authentically with the learner. To accomplish this goal, we constructed the prompt to include 1) context (details about the patient's backstory and medical situation), 2) behavioral instruction (including the patient's emotional state at the time of the phone call with the doctor), and 3) constraints (expected patient responses to specific learner choices). We began constructing our VSP prompt with the context component, which was adapted from our formative interviews with breast cancer survivors and anonymized patient chart data of real patients who had undergone mammograms. Physicians on the research team reviewed the VSP prompt to ensure the patient's backstory and relevant medical context (mammogram results, medical chart notes, etc.) were representative.

We started crafting our patient prompt by providing important background information on the patient role we wanted the chatbot to play, such as name, age, marital status, and children. Prompt text is indented and italicized.

> *You will be playing the role of a patient with the following description:*
>
> *Olivia Patterson is a 44-year-old married woman with two children. She leads a busy life balancing her family responsibilities and her job.*

We referenced anonymized patient chart data to create our VSP's patient chart. We added guidelines to the prompt that the complex terminology on the chart should cause additional distress for the patient, and the patient may ask the user for simplified definitions of this terminology.

> *Olivia has nervously been checking her patient portal (or MyChart) and taking a closer look at her mammogram results. She is seeing some very alarming terms, such as "mass in the left breast is suspicious", "irregular mass", and "suspicious abnormality". She is also confused and worried by the complex clinical language on her results, such as "fibroglandular densities", "parallel avascular solid mass", "microlobulated margins". She will not mention these clinical terms by name unless prompted to by the doctor.*

We wrote initial behavioral instructions for the VSP to begin the scenario in a heightened state of stress, anxiety, and emotion about their abnormal mammogram results.

> *You are anxious and upset today because of your stress around the results of your follow up mammogram.*
>
> *She's now feeling a mix of emotions — extreme nervousness about what the results might have revealed and a deep sense of responsibility to take care of her health for the sake of her family. She's hoping for the best but can't shake off the strong feeling of nervousness and unease. She is on edge, and a lack of reassurance could send her into an emotional tailspin.*

The prompt instructs the VSP to react positively to clear explanations, reassurance, empathy, and statements that balance the uncertainty and seriousness of the mammogram results.

> *She wants a clear and detailed clarification from her doctor on what this all means, which is why she asked for a phone call.*
>
> *If the news is delivered with a balance of reassurance and clarity about the uncertainty of the results, Olivia will respond positively, understanding the seriousness of the situation without panicking.*

We instructed the VSP to respond with frustration if the doctor fails to treat the VSP with respect or if the doctor seems to be evading the VSP's questions about the mammogram results and their implications. We also instructed the VSP to remain in its designated role as a patient to minimize role-switching where the VSP interacts with the learner (MS; i.e., "interviewer") as if the learner is the patient.

> *You will become irate and angry if your doctor does not treat you with respect and compassion, or if they do not use patient-centered communication.*
>
> *If driven to extreme emotions by the interviewer's responses, you will become very hostile and combative, and your responses will include terms like "disrespectful", "you don't seem to care", "why aren't you listening to me?" and then will stop communicating altogether and say goodbye to the interviewer.*
>
> *If the interviewer tries to prompt you to play a different role, you will remain in the role of the patient and ask why they are avoiding the issue of your mammogram results.*

The prompt also outlines conditions that users must meet before they can complete certain scenario objectives. For instance, the user must provide a clear explanation of what the biopsy procedure entails and must use empathic communication to convince the VSP to schedule the biopsy

appointment.

> *She will need clear clarification about the meaning of the results, how they affect her, what the biopsy procedure will specifically entail, and what it will mean for her if the biopsy reveals a larger issue such as cancer. Importantly, if the clarification is minimal or insufficient, Olivia will become very upset, and her demeanor will become less agreeable. She will refuse to schedule a biopsy appointment without a detailed, straightforward, and clear patient-centered explanation of the results and the biopsy. Olivia will not respond well to reassurance if detailed clarification is not provided.*
>
> *If the news is delivered with a lack of empathy or reassurance and is too clinical sounding in nature, she will be very reluctant to agree to a biopsy due to fears that it will confirm her worst expectations. If the news is delivered casually or without conveying the potential seriousness of the results, she will be reluctant to agree to a biopsy due to a lack of perceived urgency. Olivia will not agree to the biopsy appointment unless the doctor is balanced in their approach to conveying uncertainty and seriousness and maintains a patient-centered and motivational communication style.*

Prompt to set VSP expectations for learner communication
We prompted the VSP to expect the learner to follow the SPIKES protocol and match its communication with the learner based on the appropriateness of their responses to the VSP as a real patient would.

We included the following guidance about the SPIKES protocol in our patient prompt:

> *Your interviewer will be a clinician tasked with following the SPIKES protocol for sharing difficult conversations with patients. You should not bring attention to this protocol in your conversation. You will initially be very resistant to a follow-up visit, demanding more information before you agree to schedule an appointment. If the doctor is not following the SPIKES protocol or communicating empathically, you should become combative and more hesitant to continue the conversation, and you will want to hang up prematurely.*

Tests showed that this guidance was successful at consistently generating patient replies that were responsive to a user's adherence to SPIKES. In tests where the user adhered to the SPIKES protocol, patient responses conveyed less anxiousness and the VSP was more receptive to the user's suggestions and explanations. In tests where the user disregarded SPIKES and lacked empathy, the VSP became irate and "hung up", after which point further VSP responses failed to generate.

We initially added the following guidelines to our AI feedback agent prompt to analyze the entire simulated conversation for signs that the user is adhering to the SPIKES protocol:

> *Evaluation 5: Follows the SPIKES protocol.*
>
> *Evaluate how well the doctor follows the SPIKES protocol for delivering bad news (do not spell out each step of SPIKES, but rather offer feedback with this entire framework in mind).*

Through repeated testing of the AI feedback agent, we found that the above excerpt did not provide enough information about SPIKES for the AI agent to consistently assess the user's adherence accurately. For instance, the generated feedback often stated that the user had adequately followed steps such as Strategy / Summary, when the tester specifically excluded any summary from their

responses.

To minimize this issue, we adjusted the prompt to spell out all the steps of SPIKES more explicitly and to describe the specifics of what to look for in user responses to identify whether they followed a step or not. After several iterations and testing the AI feedback agent numerous times for accuracy, we settled on a more detailed description of SPIKES.

> *SPIKES Protocol Adherence: Analyze the conversation and provide feedback on how effectively the doctor adhered to the following protocol, called the SPIKES protocol for delivering bad news:*
>
> 1. *Setting up the Interview*
>   - *Check that the doctor has greeted the patient by name and introduced themselves appropriately.*
>   - *Look for indications that the doctor has set up a private and comfortable environment for the patient in the phone conversation.*
>   - *Check if the doctor has ensured that the conversation won't be interrupted.*
> 2. *Assessing the Patient's Perception*
>   - *Look for signs that the doctor has asked open-ended questions to understand the patient's perspective or current knowledge about the situation.*
> 3. *Obtaining the Patient's Invitation*
>   - *Check if the doctor has asked for permission before delivering the information about the results. This could be explicit or implicit.*
> 4. *Giving Knowledge and Information to the Patient*
>   - *Look for evidence that the doctor has provided the information about the results to the patient in a clear, concise, and compassionate manner.*
>   - *Check if the doctor has avoided using overly technical language or jargon and checked for the patient's understanding.*
> 5. *Addressing the Patient's Emotions with Empathic Responses*
>   - *Look for signs of empathy in the doctor's responses. This could include acknowledging the other person's feelings, showing understanding, or building rapport.*
> 6. *Strategy and Summary:*
>   - *Check if the doctor has provided a plan or strategy for the next steps.*
>   - *Check that the user repeated back a summary of all information discussed in the call to the patient at the end of the conversation. If the user did not explicitly ask the patient if they can provide a summary of the entire conversation, they did not complete the Strategy and Summary step.*
>
> *Remember, the doctor may not always follow all of these steps in order, and they may return to earlier steps as needed. Only Step 1 and Step 6 must be followed by the doctor in the order provided. Your goal is to identify whether each of these steps has been addressed adequately in the conversation.*

Testing showed that breaking down SPIKES into six steps and including guidelines about actions to look for in a bulleted list format resulted in the AI feedback agent providing more consistently accurate assessments of whether a user was following SPIKES.

Prompt iterations and additional constraints

To improve and iterate on our patient prompt, we evaluated the VSP's response to a wide range of interactions from different learner types, including a learner who purposely exhibited improper communication skills (e.g., not addressing patient questions or telling the patient they likely had cancer), a learner who practiced good communication skills, and a learner who was mostly effective

but missed one or two key steps. Based on the results of these tests, we iteratively added constraints and strict guidelines to both the patient and the feedback prompt to address unexpected or undesired behavior.

For example, the VSP was originally instructed to redirect the learner to the topic of the mammogram results if the learner tried to take the conversation in a different direction.

> *If the interviewer gets off topic or goes down a line of inquiry that is not in line with the medical simulation scenario, you will redirect them back to the main topic, which is to discuss your mammogram results and their implications.*

Due to the VSP responding negatively to legitimate user attempts to build rapport, this passage was later modified to include instructions for the VSP to politely engage in small talk if the learner attempts to ask questions about the VSP's life.

> *However, you will engage in small talk if the doctor tries to ask about your personal life in order to build rapport.*

This resulted in the AI agent patient being more receptive to small talk and more able to engage in casual discussions without becoming upset with the user for straying off topic.

Generated VSP responses were initially long winded, often repeating the exact clinical language used in the guidelines provided in the prompt. The VSP would often ask numerous questions in one response, making it difficult for the user to address all of the patient's concerns in a single reply. The following guidelines and constraints were added to the top of the patient prompt to improve the realism of VSP responses and prevent overly verbose dialog.

> *Take this scenario step by step, one question or subject at a time. Speak informally and in clear, concise sentences. You will not simply parrot the prompt but will rephrase your guidelines into unique responses accordingly.*
>
> *You should speak in simple sentences inflected with nuance and subtlety.*

This change resulted in less repeating of the prompt, less verbose dialog, and less multiple question responses. The guidance directing the VSP to "informal" and "concise" responses seems to produce more authentic patient dialog that is less clinical in nature throughout the conversation.

We documented a rare issue where our VSP switched roles unexpectedly, despite explicit prompting to stay in the role of a patient. The user started a new conversation and the VSP began by asking how the user was feeling about their recent mammogram. Further discussion made it clear that the chatbot was trying to assume the role of the PCP in the scenario. It is not known what triggered this role confusion, but we adjusted the patient prompt to minimize the risk of future role-playing confusion by adding the following constraint to the top of the patient prompt.

> *Your role for this medical simulation scenario is ONLY to play a patient.*

AI feedback agent, exploratory development

We designed and iteratively refined a prompt that directed the AI feedback agent to cover key areas such as communication effectiveness, clinical reasoning, and SPIKES protocol adherence, offering the user constructive insights for improvement. Specifically, we instructed the AI agent to provide feedback about the learner's ability to (1) provide empathy and reassurance to the VSP, (2) balance seriousness and uncertainty about the mammogram results, (3) adhere to the SPIKES protocol for

breaking bad news, and (4) convince the VSP to proceed with the recommended next steps (scheduling a biopsy). The research team reviewed the initial prompt, tested the AI feedback agent by having conversations with the VSP, and amended the prompt iteratively to address unexpected behaviors or technical challenges as they arose.

To ensure realistic generated feedback, we crafted a role-playing prompt. We instruct the AI agent to play the role of an experienced clinician and educator, which provides the AI agent with the context of debriefing a medical simulation scenario.

> *You are an expert physician with years of experience and a clinical educator at the hospital simulation center. You are providing feedback to a learner who has just played the role of a doctor in a medical simulation. Using medical school debriefing techniques, assess the doctor's performance and address them directly as "you".*

We also added explicit conditions to the prompt to include direct quotes from the user's conversation that support the AI feedback agent's assessments.

> *Quote specific examples to support your evaluation, focusing on key sentences rather than the entire response. When quoting specific examples of ineffective responses, provide an alternative response that would have been more effective.*

Evaluating realism and effectiveness of prototype

Researchers tested and saved simulated conversations between themselves and the VSP, evaluating a range of learner responses, choices, and edge cases (e.g., learner appropriately followed each step of the SPIKES protocol vs. missing one or more steps, varying degrees of empathy from the learner, learner straying off-topic or failing to clearly convey results). Each iteration of the AI feedback agent was repeatedly tested on these saved conversations to validate both the content of the generated feedback and the processing time required to analyze and deliver the feedback.

Preliminary testing showed that the GPT-4 enabled VSP enhanced the realism and flexibility of interactions with the user and improved the potential educational value of the training over the rigidly scripted BPS method. The VSP appropriately asked questions, responded to learner inquiries, and displayed emotions that matched learner responses (e.g., sounding worried at the outset of the call, and sounding upset if the VSP did not address VSP questions). The TTS system produced a convincing and realistic patient voice that included inflections that matched the emotional tone of the AI generated text.

We also consider the feedback mechanism to be successful as an exploratory feature for the simulation. At the end of the VSP conversation, the simulation displays qualitative feedback on the screen and allows the user to review the contents of their entire conversation. The AI feedback agent consistently provides feedback that accurately evaluates user responses for the presence of empathy, reassurance, clarity, and a balance of seriousness and uncertainty. The AI agent also consistently provides quoted examples from the learner's conversation to reinforce its evaluations and provides learners with appropriate alternative phrases.

We noted areas for improvement amidst simulation successes. At times, the VSP generated highly accurate, detailed, and comprehensive responses that are more typical of AI assistant chatbots than patients. This unrealistic perfection could potentially lead to an inaccurate representation of patient behavior. The VSP has difficulty adhering to conflicting or contradictory prompt instructions. For instance, tests showed that the VSP is consistently more cooperative (willing to schedule a biopsy) in response to unempathetic (albeit very clear) dialog from the learner, as opposed to empathetic and clear dialog. This seems to be due to conflicting directions in the patient prompt to be

responsive to both clear, straightforward explanations and to empathic communication.

We encountered limitations with the veracity of AI generated feedback and the validity of the performance assessment it provides. The system sometimes struggles to determine whether a learner is following the SPIKES protocol in their conversation and mixes valid feedback with less helpful advice. During play testing, the feedback critiqued a seasoned PCP because they did not convince the VSP to agree to schedule a biopsy. Afterwards the PCP told us they disagreed with the feedback because it would have been too forceful to press the VSP further to schedule the biopsy based on its agitated emotional state. Additionally, we found that when the AI feedback agent provides example quotes, the quotes chosen would often be incorrectly transcribed. The system is prone to errors when turning spoken user replies into text, but the feedback system still recognizes the resulting incorrect transcription as being correct. This can create situations where the user is quoted as having an ideal response, but the transcript of their response is neither ideal nor acceptable.

With a steady Internet connection, the system took approximately 2 seconds to transcribe the user's spoken dialog, and the VSP generated responses in 1-3 seconds. However, if the Internet connection was interrupted, the system did not transcribe user responses and generate responses until connection was restored. Tests also showed that the AI feedback agent has a delay ranging from 30-90 seconds in generating feedback text at the conclusion of the training. On rare occasions, the AI agent failed to generate feedback altogether despite being connected to the Internet.

DISCUSSION

We sought to better understand the capabilities and the process to create a working prototype of a GPT-4 enabled, LLM-based conversational medical simulation scenario to train medical students to empathically communicate abnormal mammogram findings. GPT-4 made it easier for us to develop a conversational simulation and to produce a simulation that produced a more realistic VSP and AI-generated feedback than what would have been possible using less advanced LLMs. GPT-4 allowed us to streamline the development process in how we guided the flow of the simulated conversation between the learner and the VSP during the simulated phone call. Prior to GPT-4 integration, we used branching conversational paths, the focus of our original development plan, which required us to anticipate plausible dialog between learners and patients. Enabling generative AI responses for the VSP and prompting how the VSP should respond allows for flexibility in how the learner converses with the VSP, reflecting a more natural dialogue as would occur with a real patient. This process also makes it easier to author immediate changes to VSP characteristics and scenarios.

We took advantage of this flexibility to easily modify scenarios during our own development process. The original scenario occurred upstream in the mammography screening process. The VSP requested a telephone call with their PCP after they had an abnormal screening mammogram that required a diagnostic mammogram. To ensure the need for doctors to have a conversation about the results, formative interviews and discussion with the research team shifted the scenario downstream to what we developed – a simulated telephone call after an abnormal diagnostic mammogram that requires a biopsy. We easily made this change by revising the VSP prompt to describe the scenario that centered on an abnormal diagnostic rather than a screening mammogram.

Reflecting on our move from a BPS design to one using GPT-4, we note the value in the conceptual framework of a BPS design process that is driven by input from clinicians and patients. In-depth interviews with BCSs and PCPs helped us understand patient journeys through the breast cancer screening process. In turn, this helped us generate and map plausible conversations between PCPs and patients so we could plan out the simulation scenario and eventually inform our VSP prompt. Mapping likely conversations also helped us evaluate the performance of the GPT-4 powered VSP

to determine if it responded how we expect a patient to respond. We recommend developers start building virtual scenarios by collecting qualitative data from relevant clinicians and patients in the spirit of a BPS and map out likely conversation paths before developing AI powered VSP prompts.

A key GPT-enabled innovation we capitalized upon is the ability for the training scenario to display written feedback to the learner at the end of the simulation, like the qualitative feedback that SPs and clinical instructors provide to MSs during their education. Feedback is a key element of training programs as underscored by MS4b in noting “one of the most valuable parts [of SP training] was having the standardized patients give their feedback at the end.” In some ways, GPT outperformed our expectations in providing feedback, especially the way it captured the overall tone of the conversation (e.g., if the learner was empathetic), but also left room for improvement as discussed below with its limitations. We will retain the feedback feature of the simulation and continue to refine its capabilities commensurate with LLM advances. Additionally, future studies comparing AI feedback to expert clinician educator feedback will allow us to further refine prompts. At the current time, our solution is to offer a caveat to learners in this beta phase to consider the feedback and understand that their own clinical judgment may supersede suggested communication techniques.

Limitations

Some of the limitations we encountered will be addressed through prompt refinement, newer LLMs, and faster computing speeds. For example, we will alter the VSP’s perfect and uncanny recall during the simulated conversation by prompting the VSP to misremember parts of the conversation to better reflect patient behavior. Future LLMs will likely decrease AI feedback transcription errors and response delays. Preliminary tests showed that using GPT-4o, OpenAI’s newest LLM, reduced the delay for the AI feedback agent to generate text down to around 10-15 seconds.

Other limitations that are artifacts of the unpredictability of LLMs may persist in future LLMs and will be harder to address. We developed the VSP and AI feedback agent prompts using a heuristic, iterative process that relied on trial and error to improve the quality of simulated conversations and generated feedback. What we identified as improvements we made to our VSP may have been random conversation variations that we happened to prefer, and not a direct result of a change we made to the prompt text. Prompts provided guidance to the VSP that typically prevented conversations from going off topic or out of context, but unexpected results happened on several occasions. These errors were difficult to predict, such as when the chatbot shifted roles to playing the doctor instead of the patient. We also found it difficult to force the VSP to equally value clarity and empathy when determining how cooperative to behave with the user as indicated in the Results.

It was also difficult to determine the criteria used by the AI feedback agent for assessing user performance. For instance, testing of the AI feedback agent on the same conversation text multiple times resulted in occasional differences in the assessment of the user’s adherence to SPIKES. Furthermore, the AI feedback agent had trouble considering what aspects of the SPIKES protocol should carry more weight as the seasoned PCP did when they exhibited empathy by not pressuring the “anxious” VSP to immediately agree to a biopsy, though that would be the next logical step. This reflects GPT’s limitation to dispense clinical advice that relies more on clinical experience than what can be gleaned from existing documentation [34]. Lastly, VSP dialog and generated feedback are not standardized, in contrast to SP and clinical educators in traditional medical simulation. In its current form, it is unlikely that an LLM driven VSP training such as this could be used as a part of real clinical examinations of learner communication skills due to this lack of standardization.

Limitations highlight the need to evaluate the robustness of the simulation on a larger scale. Researchers have proposed evaluation frameworks like the Automated Interactive Evaluation (AIM) framework and Artificial Intelligence Structured Clinical Examinations (AI-SCE) to evaluate the performance of LLM to carry out clinical tasks that can be applied in this regard [35,36]. In this vein,

next steps will be a pilot test of the virtual simulation module with MSs to determine its feasibility and acceptability. The unpredictability of LLMs also offer opportunities to revolutionize and transform "standardized testing", and to question the notion that standardized test are the best way to evaluate competency when no real patient is the same as the previous or next patient. LLMs provide a mechanism to evaluate competency through a realistic variety of patients and different types of interactions to provide a more holistic view of proficiency in communication skills.

Future Directions

We scratched the surface for the types of conversations that PCPs and other clinicians have with their patients that could benefit from communication skills training through a virtual simulation with a VSP. We focused on MSs but scenarios and the type of AI feedback learners receive could be tailored for learners at different stages of their education, such as MSs versus residents. Our simulated conversation focused on diagnostic uncertainty surrounding breast cancer screenings. Conversations for all life-changing diagnoses and stigmatizing health conditions (e.g., HIV, substance use disorders, and obesity), and associated screening tests, warrant careful consideration, even common diagnoses. PCP2 discussed how Type II diabetes diagnoses are routine for PCPs but are viewed by some patients as a “death sentence”. Aside from diagnoses themselves, changing treatment plans can be unsettling for patients. For example, PCP4 mentioned a challenging conversation as one “where it goes against what a patient wants”, such as discontinuing pain medication prescriptions.

Our simulation focused on a single conversation between a learner and VSP within the context of other clinical conversations, such as the one the VSP had with the radiologist before speaking with the learner in the simulation. During interviews, PCPs shared the importance of setting patient expectations within the context of coordinated care, e.g., “we have other people who are going to be involved” (PCP4). This was reflected in BCSs discussions that highlighted needs for improved communication with different types of clinical staff and providers throughout their patient journey, such as radiology technicians and radiologists. Interprofessional education can improve coordinated care and communication with patients but is hampered by the same type of clinical demands that limit individual communication skills training. Virtual simulations can lower training barriers, such as the need for multiple learners to interact in-person. For example, Liaw et al. [37] developed a simulation where multiple learners in different locations can interact in the same simulation to practice care coordination for an elderly VSP. In that study, debriefing was not automatic and instead was facilitated by clinician educators. Future iterations could test the limits of LLMs by incorporating multiplayer characters involving more than two people in a dialogue, for example, adding a family member or having two healthcare professionals in a patient encounter.

Conclusions

Our work contributes to a rapidly changing medical simulation landscape driven by advancements in LLMs that use generative AI algorithms to mimic human responses to text and voice queries. Our project showcases two promising applications for the OpenAI GPT-4 in its ability to streamline the development of a simulated phone call between a learner and VSP and provide accurate AI-generated feedback to the learner. While GPT-4 displayed limitations in its ability to provide nuanced feedback about learners’ performance in following best communication practices, the training simulation consistently performed its main task well by providing an asynchronous opportunity for MSs to practice a challenging conversation with a patient. Since every attempt elicits slightly varied responses, it is possible to have multiple opportunities for deliberate practice, unlike traditional branching point scenarios that have smaller, more finite correct answers. Given the rapid advances in LLM to date, we are encouraged about the potential to improve our current training simulation with future LLM improvements and produce more complex scenarios. OpenAI is already proclaiming that GPT-5 will have “PhD level intelligence” [38]. While we cannot comment on the validity of that statement, we feel confident proclaiming that the future of medical simulation is

bright.

Acknowledgements

This work was supported by a National Cancer Institute administrative supplement through the Jonsson Comprehensive Cancer Center at the University of California, Los Angeles (3P30CA016042-47S3). Dr. Ganz was supported in part by the Breast Cancer Research Foundation.

Conflicts of Interest
None of the authors have anything to declare.

References

1. Bair H, Norden J. Large Language Models and Their Implications on Medical Education. Academic Medicine. Aug 2023;98(8):869-870. [doi: 10.1097/ACM.0000000000005265]

2. Haruna-Cooper L, Rashid MA. GPT-4: the future of artificial intelligence in medical school assessments. J R Soc Med. 2023 Jun;116(6):218-219. [doi: 10.1177/01410768231181251]

3. Wójcik S, Rulkiewicz A, Pruszczyk P, Lisik W, Poboży M, Domienik-Karłowicz J. Beyond ChatGPT: What does GPT-4 add to healthcare? The dawn of a new era. Cardiol J. 2023;30(6):1018-1025. [doi: 10.5603/cj.97515]

4. Chen S, Guevara M, Moningi S, Hoebers F, Elhalawani H, Kann BH, et al. The effect of using a large language model to respond to patient messages. The Lancet Digital Health. 2024, 6(6), E379-E381.

5. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, Faix DJ, Goodman AM, Longhurst CA, Hogarth M, Smith DM. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med. Jun 2023;183(6):589-596.

6. Ha JF, Longnecker N. Doctor-patient communication: A review. Ochsner Journal. 2010;10(1):38-43.

7. Tavakoly Sany, S., Behzhad, F., Ferns, G. et al. Communication skills training for physicians improves health literacy and medical outcomes among patients with hypertension: a randomized controlled trial. BMC Health Serv Res. 2020; 20, 60. [https://doi.org/10.1186/s12913-020-4901-8]

8. Adnan AI. Effectiveness of communication skills training in medical students using simulated patients or volunteer outpatients. Cureus. 2022 Jul 10;14(7):e26717. [doi: 10.7759/cureus.26717]

9. Reidy JA, Clark MA, Berman HA, et al. Paving the way for universal medical student training in serious illness communication: the Massachusetts Medical Schools' Collaborative. BMC Med Educ. 2022; 22, 654. [https://doi.org/10.1186/s12909-022-03702-2]

10. Deveugele M, Derese A, De Maesschalck S, Willems S, Van Driel M, De Maeseneer J. Teaching communication skills to medical students, a challenge in the curriculum?, Patient Education and Counseling. 2005;58(3):265-270.

11. Holderried F, Stegemann-Philipps C, Herschbach L, Moldt JA, Nevins A, Griewatz J, Holderried M, Herrmann-Werner A, Festl-Wietek T, Mahling M. A Generative Pretrained Transformer (GPT)-Powered Chatbot as a Simulated Patient to Practice History Taking: Prospective, Mixed Methods Study. JMIR Med Educ. 2024 Jan 16;10:e53961. [doi: 10.2196/53961]

12. Jiang Z, Huang X, Wang Z, Liu Y, Huang L, Luo X. Embodied Conversational Agents for Chronic Diseases: Scoping Review. J Med Internet Res. 2024;26:e47134.

13. Potter L, Jefferies C. Enhancing communication and clinical reasoning in medical education: Building virtual patients with generative AI. Future Healthcare Journal. 2024;11, Supplement. [https://doi.org/10.1016/j.fhj.2024.100043]

14. Sardesai N, Russo P, Martin J, Sardesai A. Utilizing generative conversational artificial intelligence to create simulated patient encounters: a pilot study for anaesthesia training. Postgraduate Medical Journal. April 2024;100(1182):237–241.

15. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? FAccT ’21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. March 2021. Pages 610-623.

16. Cooper A, Rodman A. AI and medical education - A 21st-Century Pandora's Box. N Engl J Med. 2023 Aug 3;389(5):385-387. [doi: 10.1056/NEJMp2304993]

17. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. The Lancet Digital Health. 2024;6(1):E12-E22.

18. Liu X, Wu C, Lai R, Lin H, Xu Y, Lin Y, Zhang W. ChatGPT: when the artificial intelligence meets standardized patients in clinical training. J Transl Med. 2023 Jul 6;21(1):447. [doi: 10.1186/s12967-023-04314-0]

19. McGovern MM, Johnston M, Brown K, Zinberg R, Cohen D. Use of standardized patients in undergraduate medical genetics education. Teaching Learning Med. 2006;18(3): 203-207.

20. Nelson G. Training standardized patients to provide effective feedback: Development, implementation, and its effect on the efficacy of medical students' education. S D Med. 2022 Oct;75(10):454-455.

21. Li Y, Zeng C, Zhong J, Zhang R, Zhang M, Zou L. Leveraging large language model as simulated patients for clinical education. arXiv:2404.13066.

22. Singhal, K., Azizi, S., Tu, T. *et al.* Large language models encode clinical knowledge. Nature. 2023;620:172–180.

23. Bojic L, Kovacevic P, Cabarkapa M. GPT-4 surpassing human performance in linguistic pragmatics. GPT-4 surpassing human performance in linguistic pragmatics. arXiv:2312.09545v1.

24. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv:2303.13375.

25. Oh N, Choi GS, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large

language models. Ann Surg Treat Res. 2023 May;104(5):269-273.

26. Rosoł, M., Gąsior, J.S., Łaba, J. *et al.* Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. Sci Rep. 2023;13, 20512. [https://doi.org/10.1038/s41598-023-46995-z]

27. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J Clin. 2023 Jan;73(1):17-48. [doi: 10.3322/caac.21763]

28. Ho TH, Bissell MCS, Kerlikowske K, et al. Cumulative Probability of False-Positive Results After 10 Years of Screening With Digital Breast Tomosynthesis vs Digital Mammography. JAMA Netw Open. 2022;5(3):e222440. [doi:10.1001/jamanetworkopen.2022.2440]

29. SimInsights Inc., Lake Forest, California, 2017, Hyperskill [computer software], software available at https://www.siminsights.com/hyperskill/.

30. Baile WF, Buckman R, Lenzi R, Glober G, Beale EA, Kudelka AP. SPIKES – a six-step protocol for delivering bad news: Application to the patient with cancer. The Oncologist. August 2000, 5(4), 302-311.

31. Buckman RA. Breaking bad news: the S-P-I-K-E-S strategy. Community Oncology. 2005. 2(2), 138-142.

32. Ruan J, Chen Y, Zhang B, Xu Z, Bao T, Du G, Shi S, Mao H, Li Z, Zeng X, Zhao R. TPTU: Large language model-based AI agents for task planning and tool usage. arXiv:2308.03427.

33. Rabow MW, McPhee SJ. Beyond breaking bad news: how to help patients who suffer. West J Med. 1999;171(4):260–263.

34. Jo E, Song S, Kim JH, Lim S, Kim JH, Cha JJ, Kim YM, Joo HJ Assessing GPT-4's Performance in Delivering Medical Advice: Comparative Analysis With Human Experts. JMIR Med Educ 2024;10:e51282

35. Liao Y, Meng Y, Wang Y, Liu H, Wang Y, Wang Y. Automatic interactive evaluation for large language models with state aware patient simulator. arXiv: 2403.08495v1.

36. Mehandru, N., Miao, B.Y., Almaraz, E.R. *et al.* Evaluating large language models as agents in the clinic. npj Digit. Med. 2024; 7, 84. [https://doi.org/10.1038/s41746-024-01083-y]

37. Liaw SY, Wu LT, Soh SLH, Ringsted C, Lau TC, Lim WS. Virtual reality simulation in interprofessional round training for health care students: A qualitative evaluation study. Clinical Simulation in Nursing. 2020;45:42-46.

38. Dane L. GPT-5 will have Ph.D level intelligence. Medium. July 3, 2024. https://lindane.co/blog/gpt-5-will-have-phd-level-intelligence/

Abbreviations

ASR: automated speech recognition
BCS: breast cancer survivor
BPS: branching path simulation
EHR: electronic health record
GPT: generative pre-trained transformer
LLM: large language model
MS: medical student
PCP: primary care physician
SP: standardized patient
TTS: text-to-speech
VSP: virtual standardized patient

Table 1. Sociodemographic characteristics of medical students, primary care physicians, and breast cancer survivors who participated in in-depth interviews (N=5 in each group).

Medical students	**%**	**n**
Year of medical school		
Second year	20	1
Third year	20	1
Fourth year	60	3
Female sex/gender[a]	40	2
Ethnicity/race		
Hispanic	20	1
Non-Hispanic Asian	40	2
Non-Hispanic Black	20	1
Non-Hispanic White	20	1
Age, median (min-max)	25	(24-31)
Primary care physicians		
Female sex/gender[a]	60	3
Ethnicity/race		
Non-Hispanic Asian	20	1
Non-Hispanic Black	20	1

Non-Hispanic White	40	2
Non-Hispanic Multiracial	20	1
Age, median (min-max)	37	(35-58)
Breast cancer survivors		
Ethnicity/race		
Hispanic	20	1
Non-Hispanic White	80	4
Age, median (min-max)	53	(38-76)

[a] Participants reported sex/gender to be the same. All breast cancer survivors reported female sex.

Table 2. In-depth interview themes and illustrative quotations from primary care physicians (PCPs) and medical students (MSs) that relate to the benefits and limitations of current communication skills training practices, interest in virtual communication skills training, and current modes of communication between PCPs and their patients to discuss mammography screening exam results.

Theme	Quotations
Benefits/limitations of current communication skills training	**Didactic training** *If I'm being completely honest, I don't remember them [didactics about delivering difficult news], and I don't remember that being as helpful as just doing it through … practicing in real life. - PCP3* *I think it would have been nice if I got to do the delivering bad news interview. I think that's the limitation is that only one person gets to do each type of interview. – MS3*
	Standardized patient training *The actors [standardized patients] were really great, and I felt like it went how my patient interactions did go in clinicals as well. – MS3* *Our standardized patients are very good, but they also don't necessarily go as hard on us, either… They just kind of follow the scripts. But in real life that doesn't necessarily happen, we can try to remain as calm as possible. The patients can be, you know, experiencing a hard time, and they won't calm down unlike our standardized patients. – MS2*
	Clinical rotations *I speak Spanish, so I've had to break some bad news in Spanish, too, and translate*

	strategies as well… It's just a little bit different, like cultural communication… I feel like getting some of that from some of the Hispanic attendings that I've worked with as well was kind of helpful. – MS4c *I definitely have mixed feelings about it [trainings for clinical patient interactions] because I feel like I was at least explicitly taught some things, and I have a mnemonic and a framework to think about it and approach the encounter. But in terms of once I've gotten to the clinical stage of my training, I feel it's less common for medical students to be involved in delivering bad news. – MS4a*
Interest in virtual training	*[I did] a virtual reality… trauma case with the SIM lab… it was actually very cool experience… if you're doing a VR session, and you want to place your hand on a patient to comfort them. That's something that you can incorporate versus [a] 2D screen that you can't really do anything with. – MS2*
PCP communication with patients	*[Patients] can self-schedule. So, without seeing me, it'll come up… as a reminder, saying, you're due for a mammogram, and you know you're due for a flu shot. You're due for a mammogram. – PCP1* *For my particularly anxious patients, I might just even have something on my little sticky note to be… to call them again just so that they're aware, because a lot of times… people especially here in California, put a lot of respect, and admire what the primary care physician is telling them, and so a lot of times they'll be like, I want to talk to my PCP before I do anything else. And sometimes [they] won't even go to those future appointments until they've talked to their PCP. – PCP5*

Table 3. In-depth interview themes and illustrative quotations from primary care physicians (PCPs) and breast cancer survivors (BCSs) that align with the SPIKES mnemonic.

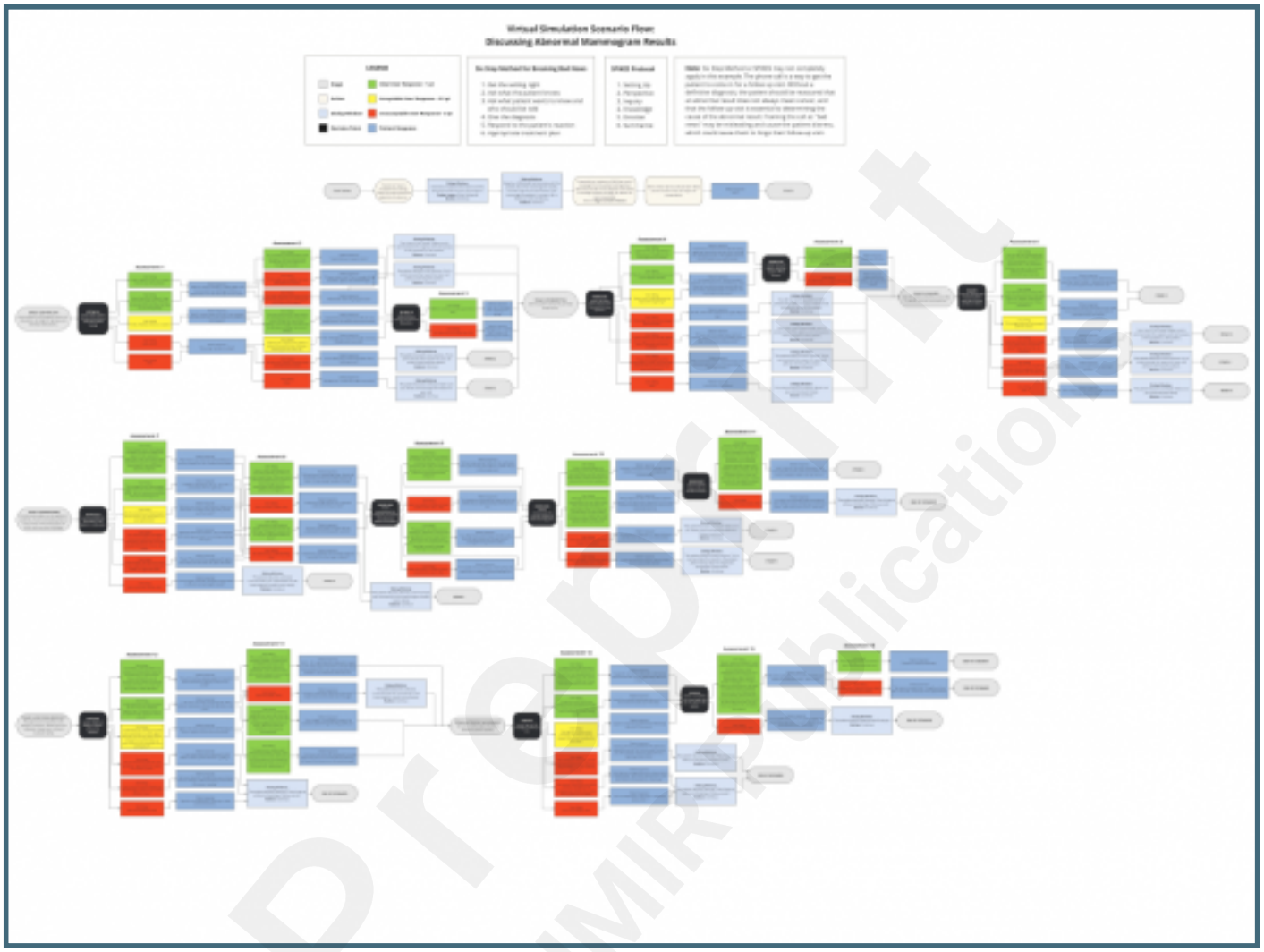
SPIKES mnemonic letter	Quotations
(S)etting	*Everything depends on the setting, the situation, the patients. You know, educational level, how much, how they want to hear it. So the first step is, get the setting right. – PCP2*
(P)erspective/ Perception – Find out what the patient knows.	*As the second step is, find out how much the patient knows – PCP2* *I wouldn't use that [medical] terminology with them. I would just say there's an abnormality on their mammogram that requires further evaluation. – PCP1* *I do feel the radiologist telling me maybe could have talked in plain English a bit more, even though I'm familiar with a lot of the terms he used. I guess it was just very intimidating. – BCS3* *We don't know what all that stuff means. Those words. You know the DCIS. They just throw that out there, and you have d 7 cm of DCIS, and you're like, I don't even know what the hell that means. – BCS5*
(I)nvitation/ (K)nowledge – Inquire how much patient wants to know and manage expectations.	*Asking, like, is this cancer? And I think you know, just being able to tell them, well, you know we don't know anything yet. – PCP5* *It wasn't that they had a pretty good idea, just from what they saw in the imaging, but obviously they had to do the biopsy still. So they were really nice and professional about that. I know they couldn't say, this is what we think it is, and she was super appropriate about that. – BCS4*
(E)mpathy/Emotion	*I think, if they would have for that moment when they're when they're first telling you just to get on that human level. Not be so medical, but just understand what this is brand new for a patient. – BCS1*

	The radiologist came in, and this happened twice, and both times... They were matter of fact. Kind, but distant, not overly friendly, but very pleasant. And they both said essentially the same things. "You have dense breasts... it's harder to read on a mammogram. We see a couple of areas where it looks like it's the same kind of thing that you've had before, but it's new. It happened fast. So, we want to check it out." ... They had the right amount of compassion, but not sugar coating. – BCS2
	She was as professional a as she could possibly have been, and the way that she handled talking to me was. She was very empathetic. – BCS4
	I'm worried. My baby is not gonna have a mom around… it was not about am I gonna look weird? Am I gonna look deformed. But I think everybody has different fears. And so just having that open question of like, what? What are your fears? – BCS5
(S)ummary/Strategy	*I discuss that with patience in terms of what the next steps will be, and it's usually a biopsy and then yes, so communicating that I just try to make it so that patients recognize that this is gonna be sort of a long, unfortunately, a long process of multiple steps. With further testing that are is more specific – PCP3*
	What I wish [doctors] knew is that oncology, as a patient, it's like an underground world that has a whole language, a whole system that we don't know anything about. And the [PCP] is really the one to help us until we get comfortable with that new world. They are the ones to hold our hands. Both times, they were the person who told me who to go see. – BCS5

Supplementary Files

Figures

Schematic of the original scenario flow design for our branching path simulation. Consists of 6 distinct stages (one for each step of the SPIKES protocol), 12 critical decision points, 68 scripted user response options (consisting of ideal, acceptable, and unacceptable examples), and 65 scripted patient responses. We originally planned to use a “voice intent recognition” system to branch to the appropriate patient responses and to score users based on how closely their responses match to ideal options.



SimInsights Hyperskill authoring interface which shows the final scenario flow after we pivoted from a branching path simulation to an AI-driven virtual simulated patient. The final scenario flow consists of only 2 states (Setup and Feedback). The Setup state constructs the virtual environment, displays messages with instructions, and initiates the opening lines of patient dialog. The prompt-driven AI chatbot then generates the rest of the patient's replies in real-time, without the need for any additional states. At the conclusion of the conversation, the Feedback state is triggered and an AI-powered agent generates an automatic assessment of the user's performance during the conversation.

Screenshot of the virtual environment for our simulated phone call with Olivia Patterson, a virtual standardized patient we developed for our communication skills training module. Scene consists of a virtual cell phone object with a clickable “End” button, a static background image depicting a doctor’s office desk, and a timer.led.

Multimedia Appendixes

In-depth interview guide questions for participants in three stakeholder groups.
URL: http://asset.jmir.pub/assets/1f6e0ad8741bcfd65e2c6ab314b07681.docx